Opinion Extraction, Summarization and Tracking in News and Blog Corpora

Lun-Wei Ku, Yu-Ting Liang and Hsin-Hsi Chen

Department of Computer Science and Information Engineering
National Taiwan University
Taipei, Taiwan
{lwku, eagan}@nlg.csie.ntu.edu.tw; hhchen@csie.ntu.edu.tw

Abstract

Humans like to express their opinions and are eager to know others' opinions. Automatically mining and organizing opinions from heterogeneous information sources are very useful for individuals, organizations and even governments. Opinion extraction, opinion summarization and opinion tracking are three important techniques for understanding opinions. Opinion extraction mines opinions at word, sentence and document levels from articles. Opinion summarization summarizes opinions of articles by telling sentiment polarities, degree and the correlated events. In this paper, both news and web blog articles are investigated. TREC, NTCIR and articles collected from web blogs serve as the information sources for opinion extraction. Documents related to the issue of animal cloning are selected as the experimental materials. Algorithms for opinion extraction at word, sentence and document level are proposed. The issue of relevant sentence selection is discussed, and then topical and opinionated information are summarized. Opinion summarizations are visualized by representative sentences. Text-based summaries in different languages, and from different sources, are compared. Finally, an opinionated curve showing supportive and non-supportive degree along the timeline is illustrated by an opinion tracking system.

Introduction

Watching specific information sources and summarizing the newly discovered opinions are important for governments to improve their services and for companies to improve their products (Dave *et al*., 2003 and Morinaga *et al*., 2002). Opinion extraction identifying components which express opinions is fundamental for summarization, tracking, and so on (Ku, Li, Wu and Chen, 2005). At document level, Wiebe, Wilson and Bell (2001) recognized opinionated documents. Pang, Lee, and Vaithyanathan (2002) classified documents by overall sentiments instead of topics. Dave's (2003) and Hu's (2004) researches focus on extracting opinions of reviews. However, a document consists of various opinions. Riloff and Wiebe (2003) distinguish subjective sentences from objective ones. Kim and Hovy (2004) propose a sentiment classifier for English words and sentences, which utilizes thesauri. However, template-based approach needs a professionally annotated corpus for learning, and words in thesauri are not always consistent in sentiment.

Hu and Liu (2004) proposed an opinion summarization of products, categorized by the opinion polarity. Liu, Hu and Cheng (2005) then illustrated an opinion summarization of bar graph style, categorized by product features. Nevertheless, they are both domain-specific. Wiebe *et al*. (2002) proposed a method for opinion summarization by analyzing the relationships among basic opinionated units within a document. Extracting opinions on products (Hu and Liu, 2004) is different from that on news or writings. For these kinds of articles, major topic detection is critical to expel non-relevant sentences (Ku, Li, Wu and Chen, 2005) and single document summarization is not enough.

Pang, Lee and Vaithyanathan (2002) showed that machine learning approaches on sentiment classification do not perform as well as that on traditional topic-based categorization at document level. Information extraction technologies (Cardie *et al*., 2004) have also been explored. A statistical model is used for sentiment words too, but the experiment material is not described in detail (Takamura *et al*., 2005). The results for various metrics and heuristics also depend on the testing situations.

News and blog articles are two important sources of opinions. The writing of the former is comparatively formal to that of the latter because blog articles expressing personal opinions of the writers are often written in a casual style. Because no queries are posed beforehand, detecting opinions is similar to the task of topic detection at sentence level. Besides distinguishing between positive and negative opinions, identifying which events correlated with which opinions are also important. This paper proposes a major topic detection mechanism to capture main concepts embedded implicitly in a relevant document set. Opinion summarization further retrieves all the relevant sentences related to the major topic from the

Copyright © 2006 American Association for Artificial Intelligence (www.aaai.org). All rights reserved.

document set, determines the opinion polarity of each relevant sentence, and finally summarizes positive sentences and negative sentences. Summaries and sentiment scores are finally used by the opinion tracking system. We will employ documents of different sources and in different languages to demonstrate the performance of opinion extraction, summarization and tracking.

Corpus Description

Three sources of information are collected for the experiments: TREC[1] corpus (Text REtrieval Conference), NTCIR[2] corpus and articles from web blogs. TREC corpus is in English, while the other two are in Chinese. Two Chinese materials are annotated for the inter-annotator agreement analysis and the experiment of opinion extraction. All of them are then used in opinion summarization. Opinion summaries about "animal cloning" of these three sources are used as illustrations.

Data Acquisition

The first corpus used is the test bed of novelty track in TREC 2003 (Soboroff and Harman, 2003). There are 50 document sets in 2003 TREC novelty corpus, and each set contains 25 documents. All documents in the same set are relevant. In TREC corpus, set 2 ("clone Dolly sheep") is taken as an example. It discussed the feasibility of the gene cloning and the perspectives of the authority.

The second corpus is NTCIR. Chen and Chen (2001) developed a test collection CIRB010 for Chinese information retrieval in NTCIR 2. The test collection consists of 50 topics and 6 of them are opinionated topics. Total 192 documents relevant to the six topics are chosen as training data in this paper. Documents of an additional topic "animal cloning" of NTCIR 3 are selected from CIRB011 and CIRB020 document collections and used for testing.

Blog is a new rising community for expressing opinions. To investigate the opinions expressed in blogs, we retrieve documents from blog portals by the query "animal cloning". The numbers of documents relevant to "animal cloning" in three different information sources are listed in Table 1.

Source	TREC	NTCIR	BLOG
Quantity	25	17	20

Table 1. Numbers of documents for opinion summarization

Annotations

To build up training and testing sets for Chinese opinion extraction, opinion tags at word, sentence and document levels are annotated by 3 annotators. We adopt the tagging format specified in the paper (Ku, Wu, Li and Chen, 2005). There are four possible values – say, positive, neutral, negative and non-sentiment, for the opinion tags at three levels. NTCIR news and web blog articles are annotated for this work.

Inter-annotator Agreement

At first, the agreement of annotations is evaluated. The agreements of tag values at word, sentence and document levels are listed in Tables 2, 3 and 4, respectively.

Annotators	A vs. B	B vs. C	C vs. A	Ave
Percentage	78.64%	60.74%	66.47%	68.62%
All agree	54.06%			

Table 2. Agreement of annotators at word level

Annotators	A vs. B	B vs. C	C vs. A	Ave
Percentage	73.06%	68.52%	59.67%	67.11%
All agree	52.19%			

Table 3. Agreement of annotators at sentence level

Annotators	A vs. B	B vs. C	C vs. A	Ave
Percentage	73.57%	68.86%	60.44%	67.62%
All agree	52.86%			

Table 4. Agreement of annotators at document level

Agreements of data from news and blogs are listed in Table 5 for comparison.

SOURCE	NTCIR		BLOG	
Level	Sentence	Document	Sentence	Document
Average agreements of two annotators	53.33%	41.18%	73.85%	64.71%
All agree	33.33%	17.65%	61.40%	41.18%

Table 5. Agreements of annotations of data from two different sources

Table 5 shows that tagging news articles has lower agreement rates than tagging web blogs. This is because blog articles may use simpler words and are easier to understand by human annotators than news articles.

From the analyses of inter-annotator agreement, we find that the agreement drops fast when the number of annotators increases. It is less possible to have consistent annotations when more annotators are involved. Here we adopt voting to create the gold standard. The majority of annotation is taken as the gold standard for evaluation. If the annotations of one instance are all different, this instance is dropped. A total of 3 documents, 18 sentences but 0 words are dropped. According to this criterion, Table 6 summarizes the statistics of the annotated testing data.

1 http://trec.nist.gov/
2 http://research.nii.ac.jp/ntcir/index-en.html

	Positive	Neutral	Negative	Non-opinionated	Total
Word	256	27	243	312	838
Sentence	48	3	93	432	576
Document	7	2	11	14	34

Table 6. Summary of testing data

Tables 7-9 show the annotation results of three annotators comparing to the gold standard (i.e., the majority). On average, an annotator can "monitor" the opinions of the whole to around 80.14%. This value can be considered as a reference when we evaluate the performance of algorithms. The decision of opinion polarities depends much on human perspectives. Therefore, the information entropy of testing data should also be taken into consideration, when comparing system performance.

Annotators	A	B	C	Average
Recall	94.29%	96.58%	52.28%	81.05%
Precision	80.51%	88.87%	73.17%	80.85%
f-measure	86.86%	92.56%	60.99%	80.14%

Table 7. Annotators' performance referring to gold standard at word level

Annotators	A	B	C	Average
Recall	94.44%	38.89%	90.97%	74.77%
Precision	71.20%	74.67%	50.19%	65.35%
f-measure	81.19%	51.14%	64.69%	65.67%

Table 8. Annotators' performance referring to gold standard at sentence level

Annotators	A	B	C	Average
Recall	100%	50%	85%	78.33%
Precision	71.43%	71.43%	65.38%	69.41%
f-measure	83.33%	58.82%	73.91%	72.02%

Table 9. Annotators' performance referring to gold standard at document level

The aim of this research is to simulate the opinions of the mass. Instinctively, a statistical model would be a good choice to detect opinionated contents. In the following section, a statistical algorithm is proposed for opinion extraction.

Opinion Extraction

The goal of opinion extraction is to detect where in documents opinions are embedded. Opinions are hidden in words, sentences and documents. An opinion sentence is the smallest complete semantic unit from which opinions can be extracted. The sentiment words, the opinion holders, and the contextual information should be considered as clues when extracting opinion sentences and determining their tendencies. Therefore, the extraction algorithm is built bottom up by detecting sentiment words at first, then identifying the opinion polarities of sentences and finally documents afterwards.

We postulate that the opinion of the whole is a function of the opinions of the parts. That is, a summary report is a function of all relevant opinionated documents, the opinion of a document is a function of all the supportive/non-supportive sentences, and the degree of a supportive/non-supportive sentence is a function of an opinion holder together with sentiment words. Opinion scores of words, which represent their sentiment degrees and polarities, are determined by the proposed formulas.

Algorithm

[Word Level]
Sentiment words are employed to compute the tendency of a sentence, and then a document. To detect sentiment words in Chinese documents, a Chinese sentiment dictionary is indispensable. However, a small dictionary may suffer from the problem of coverage. We develop a method to learn sentiment words and their strengths from multiple resources.

First we collect two sets of sentiment words, including General Inquirer[1] (abbreviated as GI) and Chinese Network Sentiment Dictionary[2] (abbreviated as CNSD). The former is in English and we translate those words into Chinese. The latter, whose sentiment words are collected from the Internet, is in Chinese. Table 10 shows the statistics of the revised dictionaries. Words from these two resources form the "seed vocabulary" in our dictionary.

Dictionary	Positive	Negative
GI	2,333	5,830
CNSD	431	1,948
Total	2,764	7,778

Table 10. Qualified seeds

Then, we enlarge the seed vocabulary by consulting two thesauri, including tong2yi4ci2ci2lin2 (abbreviated as Cilin) (Mei et al. 1982) and the Academia Sinica Bilingual Ontological Wordnet[3] (abbreviated as BOW). Cilin is composed of 12 large categories, 94 middle categories, 1,428 small categories, and 3,925 word clusters. BOW is a Chinese thesaurus with a similar structure as WordNet[4]. However, words in the same clusters may not always have the same opinion tendency. For example, 「寬恕」 (forgive: positive) and 「姑息」 (appease: negative) are in the same synonym set (synset). How to distinguish this polarity within the same cluster/synset is the major issue of using thesauri to expand the seed vocabulary and is addressed below.

1 http://www.wjh.harvard.edu/~inquirer/
2 http://134.208.10.186/WBB/EMOTION_KEYWORD/Atx_emtwordP.htm
3 http://bow.sinica.edu.tw/
4 http://wordnet.princeton.edu/

We postulate that the meaning of a Chinese sentiment word is a function of the composite Chinese characters. This is exactly how people read ideogram when they come to a new word. A sentiment score is then defined for a Chinese word by the following formula. This equation not only tells us the opinion tendency of an unknown word, but also suggests its strength. Moreover, using these equations, synonyms of different polarities are distinguishable while doing thesaurus expansion. We start the discussion from the definition of the formula of Chinese characters.

$$P_{c_i} = \frac{fp_{c_i}}{fp_{c_i} + fn_{c_i}} \quad (1)$$

$$N_{c_i} = \frac{fn_{c_i}}{fp_{c_i} + fn_{c_i}} \quad (2)$$

Where fp_{ci} and fn_{ci} denote the frequencies of a character c_i in the positive and negative words, respectively; n and m denote total number of unique characters in positive and negative words, respectively.

Formulas (1) and (2) utilize the percentage of a character in positive/negative words to show its sentiment tendency. However, there are more negative words than positive ones in the "seed vocabulary". Hence, the frequency of a character in a positive word may tend to be smaller than that in a negative word. That is unfair for learning, so a normalized version of Formulas (3) and (4) shown as follows is adopted.

$$P_{c_i} = \frac{fp_{c_i} / \sum_{j=1}^{n} fp_{c_j}}{fp_{c_i} / \sum_{j=1}^{n} fp_{c_j} + fn_{c_i} / \sum_{j=1}^{m} fn_{c_j}} \quad (3)$$

$$N_{c_i} = \frac{fn_{c_i} / \sum_{j=1}^{m} fn_{c_j}}{fp_{c_i} / \sum_{j=1}^{n} fp_{c_j} + fn_{c_i} / \sum_{j=1}^{m} fn_{c_j}} \quad (4)$$

Where P_{ci} and N_{ci} denote the weights of c_i as positive and negative characters, respectively. The difference of P_{ci} and N_{ci}, i.e., P_{ci} - N_{ci} in Formula (5), determines the sentiment tendency of character c_i. If it is a positive value, then this character appears more times in positive Chinese words; and vice versa. A value close to 0 means that it is not a sentiment character or it is a neutral sentiment character.

$$S_{c_i} = (P_{c_i} - N_{c_i}) \quad (5)$$

Formula (6) defines: a sentiment degree of a Chinese word w is the average of the sentiment scores of the composing characters $c_1, c_2, \ldots, c_p$.

$$S_w = \frac{1}{p} \times \sum_{j=1}^{p} S_{c_j} \quad (6)$$

If the sentiment score of a word is positive, it is likely to be a positive sentiment word, and vice versa. A word with a sentiment score close to 0 is possibly neutral or non-sentiment. Considering sentiment scores of words, sentiment words can be detected. With the sentiment words extracted, we are able to tell the opinion tendencies of sentences and documents.

[Sentence Level]

1. **For** every sentence
2. **For** every sentiment word in this sentence
3. **If** a negation operator appears before, then reverse the sentiment tendency.
4. **Decide** the opinionated tendency of this sentence by the function of sentiment words and the opinion holder as follows.

$$S_p = S_{opinion-holder} \times \sum_{j=1}^{n} S_{w_j} \quad (7)$$

Where S_p, $S_{opinion\text{-}holder}$, and S_{wj} are sentiment score of sentence p, weight of *opinion holder*, and sentiment score of word w_j, respectively, and n is the total number of sentiment words in p.

[Document level]

1. **For** every document
2. **Decide** the opinionated tendency of this document by the function of the opinionated tendencies of sentences inside as follows.

$$S_d = \sum_{j=1}^{m} S_p \quad (8)$$

Where S_d and S_p are sentiment scores of document d and sentence p, and m is the amount of evidence. If the topic is *anti* type, i.e. anti-construction of the dam, reverse the sentiment type.

Performance of Opinion Extraction

The gold standard is used to evaluate the performance of opinion extraction at word, sentence and document level. The performance is compared with two machine learning algorithms, i.e., SVM and the decision tree, at word level. C5 system is employed to generate the decision tree. For machine learning algorithms, qualified seeds are used for training (set A) and gold standard is used for testing (set B).

%	Non-normalized			Normalized		
	Verb	Noun	Average	Verb	Noun	Average
Precision	69.25	50.50	59.88	70.07	52.04	61.06
Recall	75.48	81.45	78.47	76.57	82.26	79.42
f-measure	72.23	62.35	67.29	73.18	63.75	68.47

Table 11. Performance of sentiment word mining

SVM		Testing	
		A	B
Training	A	92.08%	45.23%
	B	48.39%	56.87%

Table 12. Result matrix of SVM (Precision)

C5		Testing	
		A	B
Training	A	83.60%	36.50%
	B	0%	41.50%

Table 13. Result matrix of decision tree (Precision)

As Tables 11-13 show, the proposed sentiment word mining algorithm achieves the best average precision 61.06% of Verb and Noun while SVM achieves 46.81% (outside test, average of 45.23% and 48.39%) while C5 does even worse (precision 0% because of a small training set). Our algorithm outperforms SVM and the decision tree in sentiment word mining. This is because the semantics within a word is not enough for a machine learning classifier. In other words, machine learning methods are not suitable for word level opinion extraction. In the past, Pang *et al*. (2002) showed that machine learning methods are not good enough for opinion extraction at document level. Under our experiments, we conclude that opinion extraction is beyond a classification problem.

Source	NTCIR	WEB BLOG
Precision	34.07%	11.41%
Recall	68.13%	56.60%
f-measure	45.42%	18.99%

Table 14. Opinion extraction at sentence level

Source	NTCIR	WEB BLOG
Precision	40.00%	27.78%
Recall	54.55%	55.56%
f-measure	46.16%	37.04%

Table 15. Opinion extraction at document level

Table 14 and Table 15 show that the precision rates are low at both the sentence and document levels. This is because the current algorithm only considers opinionated relations but not relevant relations. Many sentences, which are non-relevant to the topic "animal cloning", are included for opinion judgment. The non-relevant rate is 50% and 53% for NTCIR news articles and web blog articles, respectively.

Extracting opinions only is not enough for opinion summarizations. The focus of opinions should also be considered. In the following opinion summarization section, a relevant sentence selection algorithm is introduced and applied when extracting sentences for opinion summarizations. The experimental results of opinion extraction considering relevant relations are also listed.

Opinion Summarization

Traditional summarization algorithms rely on the important facts of documents and remove the redundant information. Unlike the traditional algorithms, two factors – say, the sentiment degree and the correlated events, play the major roles of opinion summarization. The repeated opinions of the same polarity cannot be dropped because they strengthen the sentiment degree. However, the redundant reasons why they hold this position should be removed when generating opinion summaries. And needless to say, opinions must be detected first for the opinion summarization. All these reasons make the opinion summarization more challenging.

Opinion summarization aims to produce a cross-document opinionated summary. For this purpose, we need to know which sentences are opinionated and tell if they focus on a designated topic. An algorithm, which decides the relevance degree and the sentiment degree, is proposed in this section. To put emphasis on the opinionated factor, the visualization of opinion summaries is different from the traditional summaries. A text-based summary categorized by opinion polarities is also illustrated in this section. Then, a graph-based summary along time series is illustrated by an opinion tracking system.

Algorithm

Choosing representative words that can exactly present the main concepts of a relevant document set is the main work of relevant sentence retrieval. A term is considered to be representative if it appears frequently across documents or appears frequently in each document (Fukumoto and Suzuki, 2000). Such terms form the major topic of the relevant document set. How to choose the major topic is described as follows. We assign weights to each word both at document level and paragraph level. In the following formulas, ***W*** denotes weights; ***S*** is document level while ***P*** is paragraph level. ***TF*** is term frequency, and ***N*** is word count. In the subscripts, symbol *i* is the document index, symbol *j* is the paragraph index, and symbol *t* is the word index. Formulas (9) and (10) compute TF*IDF scores of term *t* in document *i* and paragraph *j*, respectively. Formulas (11) and (12) denote how frequently term *t* appears across documents and paragraphs. Formulas (13) and (14) denote how frequently term *t* appears in each document and in each paragraph.

$$W_{S_i t} = TF_{S_i t} \times \log \frac{N}{N_{S_i}} \tag{9}$$

$$W_{P_j t} = TF_{P_j t} \times \log \frac{N}{N_{P_j}} \tag{10}$$

$$Disp_{St} = \sqrt{\frac{\sum_{i=1}^{m} (W_{S_i t} - mean)^2}{m}} \times TH \tag{11}$$

$$Dev_{S_i t} = \frac{W_{S_i t} - mean}{Disp_{St}} \tag{12}$$

$$Disp_{Pt} = \sqrt{\frac{\sum_{j=1}^{n} (W_{P_j t} - mean)^2}{n}} \tag{13}$$

$$Dev_{P_j t} = \frac{W_{P_j t} - mean}{Disp_{Pt}} \tag{14}$$

A term is thought as representative if it satisfies either Formulas (15) or (16). Terms satisfying Formula (15) tend to appear in few paragraphs of many documents, while terms satisfying Formula (16) appear in many paragraphs of few documents. The score of a term, defined as the absolute value of $\mathrm{Dev}_{\mathrm{P_j t}}$ minus $\mathrm{Dev}_{\mathrm{S_i t}}$, measures how significant it is to represent the main concepts of a relevant document set.

$$Disp_{St} \le Disp_{Pt} \quad \exists S_i, \forall P_j \in S_i \; Dev_{S_i t} \le Dev_{P_j t} \quad (15)$$

$$Disp_{St} > Disp_{Pt} \quad \exists S_i, \forall P_j \in S_i \; Dev_{S_i t} > Dev_{P_j t} \quad (16)$$

Comparing with Fukumoto and Suzuki (2000), we modify the scoring function in paragraph level. All documents in the same corpus are concatenated into a bigger one, i.e., the original boundaries between documents are neglected, so that words which repeat frequently among paragraphs will be chosen. We also use threshold *TH* to control the number of representative terms in a relevant corpus. The larger the threshold *TH* is, the more the number of terms will be included. The value of this parameter is trained in the experiments. The performance of extracting relevant sentences is good (Ku *et al*., 2005).

In the opinion summarization algorithm, sentences that are relevant to the major topic and also express opinions are extracted. There are two clues for extraction, i.e., concept keywords and sentiment words. The former determines the relevance of a sentence to the topic, and the latter identifies the degree of the sentence opinion. In our experiments, concept keywords are from a predefined field (NTCIR) or automatically extracted from documents of the same topic, That is, set 2 of TREC, topic ZH037 of NTCIR, and "animal cloning" for blog articles.

The opinion-oriented sentences are extracted from topical set and their tendencies are determined. Compared to the sentence-level opinion extraction algorithm in last section, detecting the topical sentences is the first step. The overall procedure is shown as follows.

1. **For** every topical sentence
2. **For** every sentiment word in this sentence
3. **If** a negation operator appears nearby, reverse the sentiment tendency. Every sentiment word contributes its sentiment score to this sentence.
4. **Decide** the opinion tendency of a sentence by the functional composition of sentiment words, i.e., Formula (7).

Sentiment words are necessary to decide opinion polarities. If the total score of sentiment words is positive/negative, the sentence is positive/negative-oriented. Besides, we also consider opinion operators, e.g., "say", "present", "show", "suggest", *etc*. If a sentence contains such an opinion operator that follows a named entity with zero opinion score, it is regarded as a neutral opinion.

Source	NTCIR	
Level	Sentence	Document
Precision	57.80%	76.56%
Recall	67.23%	72.30%
f-measure	62.16%	74.37%

Table 16. Opinion extraction results considering concept words

As we have mentioned, major topic detection is required for opinion summarization. Table 16 shows the results of considering relevance relations together with sentiments. NTCIR corpus, with TREC style, contains concept words for each topic. These words are taken as the major topic for the opinion extraction. Sentences contain at least one concept word are considered relevant to the topic.

Obviously, the results are much better than those in Tables 14 and 15. However, in the real applications, the major topics are not available. For web blog articles, words represent the major topic must be selected automatically. The algorithm for choosing representative words is adopted and opinionated sentences are extracted again. Table 17 shows the experimental results.

Source	NTCIR	Blog
Precision	38.06%	23.48%
Recall	64.84%	50.94%
f-measure	47.97%	32.58%

Table 17. Opinion extraction results considering automatically extracted topical words

Comparing to Table 14, the precision increases after applying major topic detection algorithm. It concludes that the relevant sentence selection is an important issue in the opinion summarization.

Totally 29.67% and 72.43% of non-relevant sentences are filtered out for news and web blog articles, respectively. The performance of filtering non-relevant sentences in blog articles is better than that in news articles. The result is also consistent with the higher agreement rate of annotations in blog articles. Total 15 topical words are extracted automatically from blog articles while more, 73 topical words, are extracted from news articles. These all tell that the content of news articles diverge more than that of blog articles. However, the judgment of sentiment polarity of blog articles is not simpler (precision 38.06% vs. 23.48%).

The topical degree and the sentiment degree of each sentence are employed to generate opinion summaries. We distinguish between positive and negative documents. A document is positive if it consists of more positive-topical sentences than negative-topical ones; and vice versa. Among positive and negative documents, two types of opinion summarizations are proposed, that is, brief and detailed opinion summary. For brief summary, we pick up the document with the largest number of positive or negative sentences and use its headline to represent the overall summary of positive-topical or negative-topical sentences. For detailed summary, we list positive-topical and negative-topical sentences with higher sentiment degree. Examples of brief and detailed summaries are shown in Tables 18 and 19, respectively.

Positive	Chinese Scientists Suggest Proper Legislation for Clone Technology
Negative	UK Government Stops Funding for Sheep Cloning Team

Table 18. Brief Opinion Summary of TREC

Positive	Ahmad Rejai Al-Jundi, Assistant Secretary General of the Islamic Organization, declared earlier that the seminar would be aimed at shedding light on medical and legal aspects of the internationally controversial issue and seeking to take a stand on it.
Negative	Dolly the cloned sheep is only 3, but her genes are already showing signs of wear and she may be susceptible to premature aging and disease -- all because she was copied from a 6-year-old animal, Scottish researchers say.

Table 19. Detailed Opinion Summary of TREC

Opinion Summaries of News and Blogs

Comparing to the opinion summary of TREC set 2 (shown in Table 19), Tables 20 and 21 list the opinion summaries for NTCIR and blog articles using the same topic "animal cloning".

News and blog articles are two main sources for opinions. Different sources of articles enrich the content. Generally speaking, news documents are more objective while blog articles are usually more subjective. Besides, the opinion holders from two sources are of different social classes. The opinions extracted from news are mostly from famous people. Instead, the opinions expressed in blogs may come from a no name. Listing opinions of different sources in parallel provides views of the same public issue.

The opinion summarization algorithm proposed is language independent. With this method, opinions of different countries are visible. Moreover, this is surely the prototype of the cross lingual opinion summarization.

Positive	上述建議來自四名科學家所組成的專家小組，該小組於一月應英國政府之邀成立，就複製所衍生的法律與倫理問題提出相關建議。 (The above suggestion came from a group of four scientists. The group was formed under the request of the British government. The group was to provide advices on laws and theories concerning cloning.)
Negative	在複製羊成功的消息宣布之後，美國總統柯林頓及「生物倫理顧問委員會」斥複製人不道德，柯林頓禁止使用聯邦經費從事複製人類的實驗，並要求民間自我克制不作這種研究。 (After the announcement of the success in sheep cloning, U.S. President Clinton and National Bioethics Advisory Commission reproved human cloning as immoral. Clinton forbade using federal funds for human cloning experiments and asked the general public to refrain from doing such research.)

Table 20. Detailed Opinion Summary of NTCIR

Positive	而複製技術如果成熟，它將會是一種強大有用的工具，任何工具都可能被善用或誤用，評價一個工具不能只拿它被誤用的情境去批評它，因而禁制了它被善用的原始目的與機會，妥善的立法規範管理似乎才是較理性的作為。 (When the cloning technology reaches maturity, it will become a powerful and useful tool. Any tool can be used for a good cause or misused, so we should not blatantly criticize and dismiss a tool for its possible abuses and deprive it of its opportunity to be used in a good way. Instead, we should come up with suitable regulations for using the tool.)
Negative	有人反對複製人，因為違反了上帝的旨意。 (Some people are against cloning human beings, because it conflicts with teachings of God.)

Table 21. Detailed opinion summary of blog articles

An Opinion Tracking System

Although opinion summaries can be generated by using opinion extraction and opinion summarization algorithms, they may be distributed discretely when relevant documents are large. Like an event, we are more concerned of how opinions change over time. An opinion tracking system aims to tell how people change their opinions as time goes by. A certain quantity of articles is necessary for such analysis. Because the number of articles relevant to "animal cloning" is not large enough to track opinions in NTCIR corpus, we take the president election in the year 2000 in Taiwan as an illustrating example. Opinions towards four persons in March 2000 are shown in Figure 1.

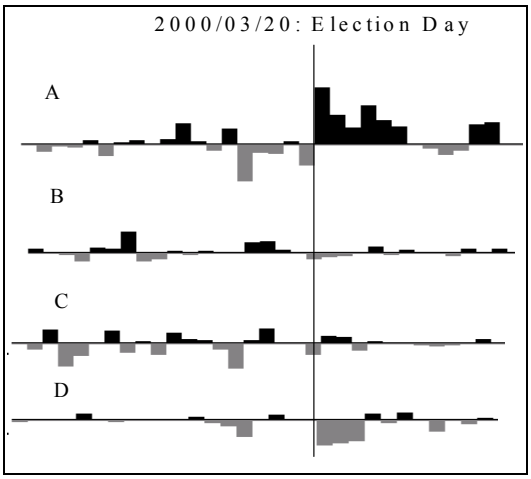

Figure 1. Opinions towards four persons

Persons A, B and C were candidates and D was the president at that time. Person A was the President elect. The trend fits the opinions in this period and the opinion summaries can tell events correlated with these opinions. This tracking system can also track opinions according to

different requests and different information sources, including news agencies and the web. Opinion trends toward one specific focus from different expressers can also be compared. This information is very useful for the government, institutes, companies, and the concerned public.

Conclusion and Future Work

This paper proposes algorithms for opinion extraction, summarization, and tracking. Different materials in different languages are experimented and compared. The nature of news and blog articles is quite different. Compared with blog articles, news articles own a larger vocabulary. On the one hand, that makes the relevant sentence retrieval harder. On the other hand, a larger vocabulary helps when deciding sentiment polarities.

The sentiment word miner mines positive and negative sentiment words and their weights on the basis of Chinese word structures. The f-measure is 73.18% and 63.75% for verbs and nouns, respectively. Experimental results also tell that machine learning methods are not suitable for sentiment word mining. Utilizing the sentiment words mined together with topical words, we achieve f-measure 62.16% at the sentence level and 74.37% at the document level. Involving topical words enhances the performance of opinion extraction.

An opinion tracking system provides not only text-based and graph-based opinion summaries, but also the trend of opinions from many information sources. Opinion summaries show the reasons for different stands people take on public issues.

Opinion holders are considered in this research. Experts or government officers have more influence when expressing opinions. However, how opinion expressers influence the sentiment degree has not yet been explored. Identifying opinion holders is very important for analyzing opinions. Properly deciding the power of opinion holders not only tells reliable sentiment degree, but also answers to the opinionated questions. Moreover, the relations between holders and their opinions are the key to solve multi-perspective problems in opinions.

Experimental resources and tools in this paper are available at http://nlg18.csie.ntu.edu.tw:8080/opinion/index.html.

Acknowledgements

Research of this paper was partially supported by National Science Council, Taiwan, under the contract NSC94-2752-E001-001-PAE.

References

Cardie, C., Wiebe, J., Wilson, T. and Litman, D. Combining low-level and summary representations of opinions for multi-perspective question answering. *Proceedings of AAAI Spring Symposium Workshop*, pages 20-27. 2004.

Chen, K.-H. and Chen, H.-H. "Cross-Language Chinese Text Retrieval in NTCIR Workshop – Towards Cross-Language Multilingual Text Retrieval." *ACM SIGIR Forum*, **35**(2), pages 12-19, 2002.

Dave, K., Lawrence, S., and Pennock, D.M. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. *WWW 2003*, pages 519-528, 2003.

Fukumoto, F. and Suzuki, Y. Event Tracking based on domain dependency. *SIGIR 2000*, pages 57-64, 2000.

Hu, Minqing and Liu, Bing. Mining and Summarizing Customer Reviews. *SIGKDD 2004*, pages 168-177, 2004.

Hu, Minqing and Liu, Bing. Mining Opinion Features in Customer Reviews. *AAAI 2004*, pages 755-760, 2004

Ku, L.-W., Li, L.-Y., Wu, T.-H. and Chen., H.-H. Major topic detection and its application to opinion summarization. *SIGIR 2005*, pages. 627-628, 2005.

Ku, L.-W., Wu, T.-H., Li, L.-Y. and Chen., H.-H. Construction of an Evaluation Corpus for Opinion Extraction. *NTCIR 2005*, pages. 513-520, 2005.

Liu, B., Hu, M. and Cheng, J. Opinion Observer: Analyzing and comparing opinions on the web. *WWW 2005*, pages 342-351. 2005.

Kim, Soo-Min and Hovy, Eduard. Determining the Sentiment of Opinions. *Coling*, pages 1367-1373, 2004.

Mei, J., Zhu, Y. Gao, Y. and Yin, H.. *tong2yi4ci2ci2lin2*. Shanghai Dictionary Press. 1982.

Morinaga, S., Yamanishi, K., Tateishi, K. and Fukushima, T. Mining product reputations on the web. *ACM SIGKDD 2002,* pages 341-349, 2002.

Pang, B., Lee, L., and Vaithyanathan, S. Thumbs up? Sentiment classification using machine learning techniques. *Proceedings of the 2002 Conference on EMNLP*, pages 79-86. 2002.

Riloff, E. and Wiebe, J. Learning extraction patterns for subjective expressions. *Proceedings of the 2003 Conference on EMNLP*, pages 105-112. 2003.

Soboroff, I. and Harman, D. Overview of the TREC 2003 novelty track. *The Twelfth Text REtrieval Conference,* National Institute of Standards and Technology, pages 38-53, 2003.

Takamura, H., Inui, T. and Okumura, M.. Extracting Semantic Orientations of Words Using Spin Model. *ACL 2005*, pages 133-140. 2005.

Wiebe, J., Wilson, T., and Bell, M. Identify collocations for recognizing opinions. *Proceedings of ACL/EACL2001 Workshop on Collocation*. 2001.

Wiebe, J., Breck, E., Buckly, C., Cardie, C., Davis, P., Fraser, B., Litman, D., Pierce, D., Riloff, E., and Wilson, T. NRRC summer workshop on multi-perspective question answering, final report. *ARDA NRRC Summer 2002 Workshop*. 2002.